IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Combined 5x2cv F Test for Comparing Supervised Classification Learning Algorithms

Ethem Alpaydın [a]

IDIAP–RR 98-04

May 1998

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Visiting scholar at IDIAP, Martigny Switzerland while this work was done. On sabbatical leave from the Department of Computer Engineering, Boğaziçi University, TR-80815 Istanbul Turkey. alpaydin@boun.edu.tr

IDIAP Research Report 98-04

Combined 5x2cv F Test for Comparing Supervised Classification Learning Algorithms

Ethem Alpaydın

May 1998

Abstract. Dietterich [1] reviews five statistical tests proposing the 5x2cv t test for determining whether there is a significant difference between the error rates of two classifiers. In our experiments, we noticed that the 5x2cv t test result may vary depending on factors that should not affect the test and we propose a variant, the combined 5x2cv F test, that combines multiple statistics to get a more robust test. Simulation results show that this combined version of the test has lower Type I error and higher power than 5x2cv proper.

1 Introduction

Given two learning algorithms and a training set, we want to test if the two algorithms construct classifiers that have the same error rate on a test example. The way we proceed is as follows: Given a labelled sample, we divide it into a training set and test set (or many such pairs) and we train the two algorithms on the training set and we test them on the test set. We define a statistic computed from the errors of the two classifiers on the test set, which if our assumption that they do have the same error rate — the null hypothesis — holds, obeys a certain distribution. We then check the probability that the statistic we compute actually has a high enough probability of being drawn from that distribution. If so we accept the hypothesis, otherwise we reject and say that the two algorithms generate classifiers of different error rates. If we reject when no difference exists, we incur a Type I error. If we accept when a difference exists, we incur a Type II error. $1 - Pr\{\text{Type II error}\}$ is called the power of the test and is the probability of detecting a difference when a difference exists.

Dietterich [1] analyzes in detail five statistical tests and concludes that two of them, McNemar test, and a new test, the 5x2cv t test, have low Type I error and reasonable power. He proposes to use McNemar test if due to high computational cost, the algorithms can be executed only once. For algorithms that can be executed ten times, he proposes to use the 5x2cv t test.

2 5x2cv Test

In the 5x2cv t test, proposed by Dietterich [1], we perform 5 replications of 2-fold cross-validation. In each replication, the dataset is divided into two equal-sized sets. $p_i^{(j)}$ is the difference between the error rates of the two classifiers on fold $j = 1, 2$ of replication $i = 1, \ldots, 5$. The average on replication i is $\overline{p}_i = (p_i^{(1)} + p_i^{(2)})/2$ and the estimated variance is $s_i^2 = (p_i^{(1)} - \overline{p}_i) + (p_i^{(2)} - \overline{p}_i)^2$.

Under the null hypothesis, $p_i^{(j)}$ is the difference of two identically distributed proportions so can be safely treated as a normal distribution with zero mean and unknown variance σ^2 [1]. Then $p_i^{(j)}/\sigma$ is unit normal. If $p_i^{(1)}$ and $p_i^{(2)}$ are independent normals, s_i^2/σ^2 is chi-square with one degree of freedom. Then

$$M = \frac{\sum_{i=1}^{5} s_i^2}{\sigma^2}$$

is chi-square with 5 degrees of freedom. If $Z \sim \mathcal{Z}$ and $X \sim \mathcal{X}_n^2$

$$T_n = \frac{Z}{\sqrt{X/n}}$$

is t-distributed with n degrees of freedom. Therefore

$$t = \frac{p_1^{(1)}}{\sqrt{M/5}} = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \tag{1}$$

is approximately t-distributed with 5 degrees of freedom [1]. We reject the hypothesis that the two classifiers have the same error rate with 95 percent confidence if t is greater than 2.571.

We note that the numerator $p_1^{(1)}$ is arbitrary and actually there are ten different values that can be placed in the numerator, i.e., $p_i^{(j)}, j = 1, 2, i = 1, \ldots, 5$, leading to ten possible statistics

$$t_i^{(j)} = \frac{p_i^{(j)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \tag{2}$$

Changing the numerator corresponds to changing the order of replications or folds and should not affect the result of the test. A first experiment is done on eight datasets to measure the effect of

Table 1: Comparison of the 5x2cv t test with its Combined version. Just changing the order of folds or replications (using a different numerator), the 5x2cv t test sometimes give different results whereas the combined version takes into account all ten statistics and averages over this variability.

	LP vs MLP	
	5x2cv $t_i^{(j)}$ rejects out of 10	Combined 5x2cv F rejects
GLASS	0	no
WINE	0	no
IRIS	2	no
THYROID	2	no
VOWEL	2	no
ODR	8	yes
DIGIT	7	yes
PEN	10	yes

changing the numerator where we compare a single layer perceptron (LP) with a multilayer perceptron with one hidden layer (MLP). ODR, DIGIT are two datasets on optical handwritten digit recognition and PEN is on pen-based handwritten digit recognition. These three datasets are available from the author. The other datasets are from the UCI repository [2].

As shown in Table 1, depending on which of the ten numerators we use, i.e., which of the ten $t_i^{(j)}, j = 1, 2, i = 1, \ldots, 5$ we calculate, the test sometimes accepts and sometimes rejects the hypothesis. That is if we change the order of folds or replications, we get different test results; this is disturbing as this order is not a function of the error rates of the algorithms and should clearly not affect the result of the test.

3 Combined 5x2cv F test

A new test that combines the results of the ten possible statistics promises to be more robust. If $p_i^{(j)}/\sigma \sim \mathcal{Z}$, then $\left(p_i^{(j)}\right)^2/\sigma^2 \sim \mathcal{X}_1^2$ and

$$N = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}\left(p_i^{(j)}\right)^2}{\sigma^2}$$

is chi-square with 10 degrees of freedom. If $X_1 \sim \mathcal{X}_n^2$ and $X_2 \sim \mathcal{X}_m^2$ then

$$\frac{X_1/n}{X_2/m} \sim F_{n,m}$$

Therefore, we have

$$f = \frac{N/10}{M/5} = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}\left(p_i^{(j)}\right)^2}{2\sum_{i=1}^{5} s_i^2} \tag{3}$$

is approximately F distributed with 10 and 5 degrees of freedom. For example we reject the hypothesis that the algorithms have the same error rate with 0.95 confidence if the statistic f is greater than 4.74. Looking at Table 1, we see that the combined version combines the ten statistics and is more robust; it is as if the combined version "takes a majority vote" over the ten possible 5x2cv t test results.

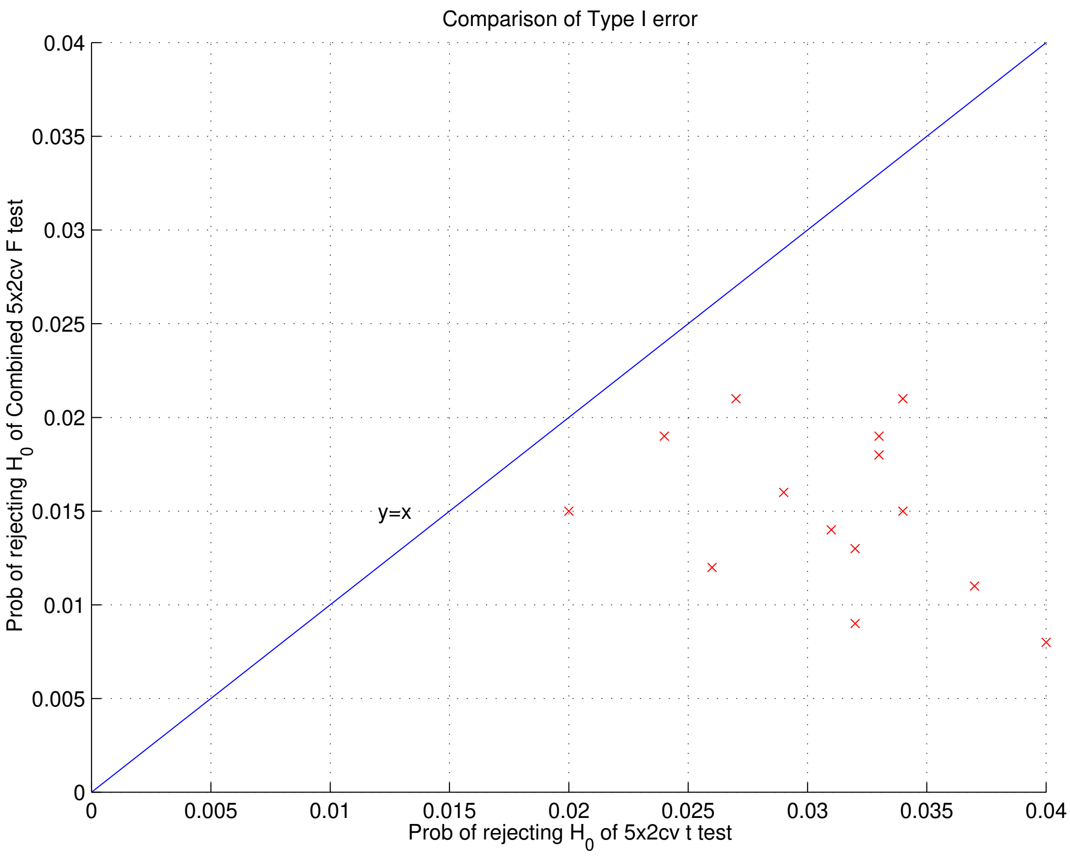

Figure 1: Type I errors of two tests are compared. All the points are under the $y = x$ line; combined test leads to lower Type I error.

4 Comparing Type I and Type II Errors

To compare the Type I error of 5x2cv with its combined version, we use two MLP with equal number of hidden units. Thus the hypothesis is true and any reject is a Type I error. On six datasets using different number of hidden units, we have designed 15 experiments of 1000 runs each. In each run, we have a 5x2cv t test result (Eq. 1) and one combined 5x2cv F result (Eq. 3). As shown in Fig. 1, the combined test has a lower probability of rejecting the hypothesis that the classifiers have the same error rate when the hypothesis is true and thus has lower Type I error. The details are given in the Appendix.

To compare the Type II error of the two tests, we take two classifiers which are different; these are a linear perceptron (LP) and a MLP with hidden units. Again on six datasets using different number of hidden units, we have designed 15 experiments of 1000 runs each where in each run, we have a 5x2cv t test result and a combined 5x2cv F result. More details are given in the Appendix.

As shown in Fig. 2, the combined test has a lower probability of rejecting the hypothesis when the two classifiers have similar error rates (lower Type II error) and has a larger probability of rejecting when they are different (higher power). The normalized difference in error rate between two classifiers is computed as

$$z = \frac{\overline{e_{lp}} - \overline{e_{mlp}}}{s_{mlp}}$$

where $\overline{e_{mlp}}, s_{mlp}$ are the average and stdev of error rate of the MLP over the test folds. Note that z is

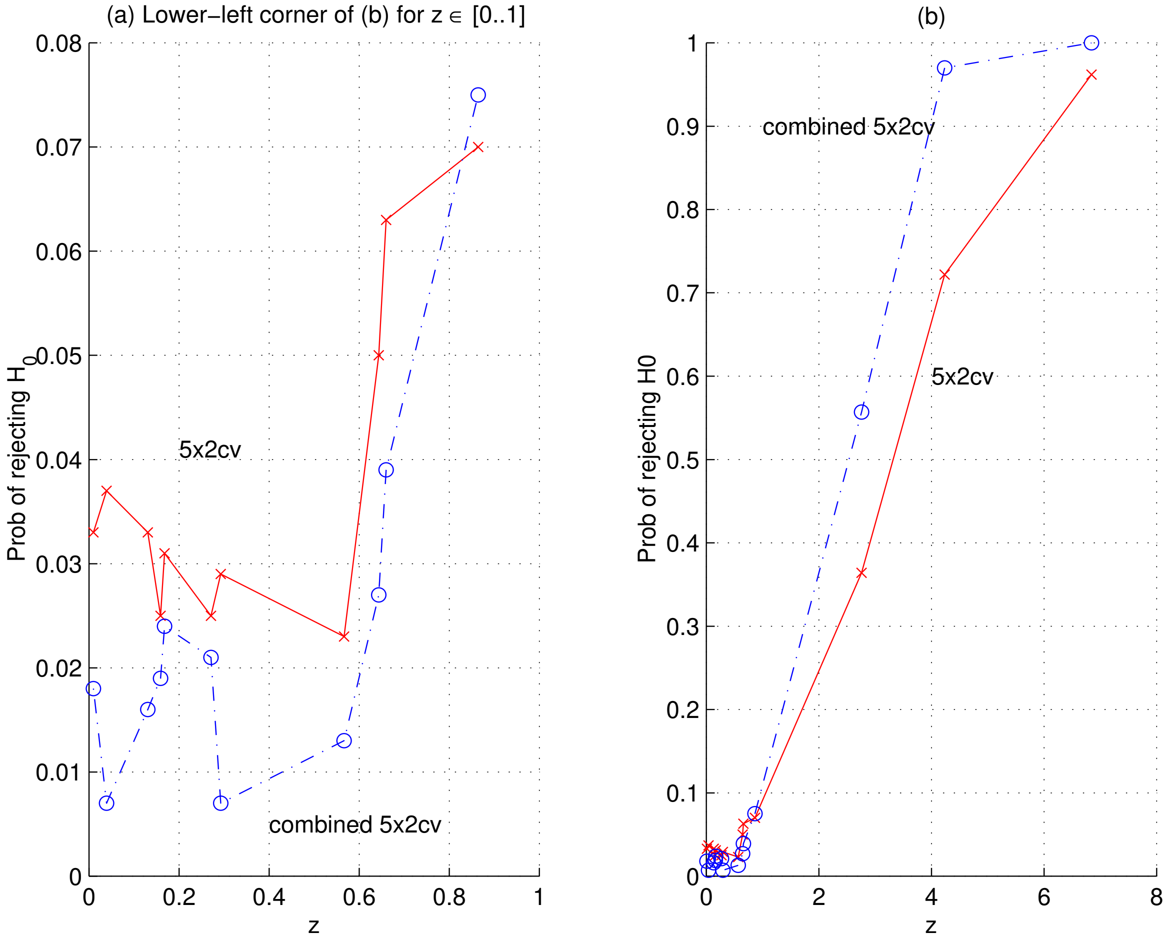

Figure 2: Type II errors of two tests are compared. (a) zooms the lower left corner of (b) for small z, the normalized distance between the error rates of the two classifiers. The combined test has a lower probability of rejecting the hypothesis when the two classifiers have similar error rates and larger when they are different.

Table 2: Average and standard deviations of error rates on test folds of a linear perceptron and multilayer perceptrons with different number of hidden units (given before ':').

	LP	MLP	MLP	MLP
IRIS	3.75, 2.05	3: 3.85, 2.57	10: 3.18, 1.95	20: 2.77, 1.73
WINE	2.84, 1.66	3: 2.86, 2.02	10: 2.57, 1.61	20: 2.63, 1.61
GLASS	38.66, 4.03	5: 37.52, 4.21	10: 35.81, 4.32	20: 35.04, 4.19
VOWEL	38.70, 2.48	5: 36.86, 2.86	10: 27.69, 2.60	20: 22.48, 2.37
ODR	5.31, 1.08	10: 5.14, 1.07	20: 3.16, 0.78	
THYROID	4.61, 0.38	10: 4.26, 0.34		

an approximate measure for what we are trying to test, i.e, whether the two classifiers have different error rates.

Small difference in error rate implies that the different algorithms construct two similar classifiers with similar error rates thus the hypothesis should not be rejected. For large difference, the classifiers have different error rates and the hypothesis should be rejected.

5 Conclusions

The combined version of the 5x2cv t test, named the combined 5x2cv F test, that averages over the variability due to replication and fold order, as the simulation results indicate, has lower Type I error and higher power than the 5x2cv t test proper.

Appendix

On six datasets we trained a one-layer linear perceptron (LP) and multilayer perceptrons (MLP) with different number of hidden units to check for Type I and Type II errors. The average and standard deviation of test error rates for LP and MLP are given in Table 2. Reject probabilities with the 5x2cv t test and the combined 5x2cv F test are given in Table 3. The probabilities are computed as proportions of rejects over 1000 runs.

Acknowledgments

Thanks to Eddy Mayoraz, Frédéric Gobry and Miguel Moreira for stimulating discussions.

References

[1] T. G. Dietterich (1998) "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, to appear.

[2] C. J. Merz, P. M. Murphy (1998). UCI Repository of Machine Learning Databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Table 3: Probabilities of rejecting the null hypothesis, i.e., proportions of times the corresponding test rejected in 1000 trials. When comparing two MLPs with equal number of hidden units, any reject is a Type I error and when comparing an LP with an MLP, if their accuracies are different, any reject is lower Type II error and implies higher power.

		MLP vs MLP (Type I error)		LP vs MLP (Type II error)	
	hid	5x2cv	Combined 5x2cv	5x2cv	Combined 5x2cv
IRIS	3	0.032	0.009	0.037	0.007
	10	0.040	0.008	0.029	0.007
	20	0.029	0.016	0.023	0.013
WINE	3	0.037	0.011	0.033	0.018
	10	0.032	0.013	0.031	0.024
	20	0.047	0.016	0.033	0.016
GLASS	5	0.034	0.021	0.025	0.021
	10	0.026	0.012	0.063	0.039
	20	0.047	0.015	0.070	0.075
VOWEL	5	0.033	0.018	0.050	0.027
	10	0.027	0.021	0.722	0.970
	20	0.034	0.015	0.962	1.000
ODR	10	0.033	0.019	0.025	0.019
	20	0.024	0.019	0.364	0.557
THYROID	10	0.031	0.014	0.041	0.031